

Nonparametric identification of causal effects with confounders subject to instrumental missingness

Shu Yang*, Linbo Wang†, Peng Ding‡

Abstract

We consider causal inference from observational studies when confounders have missing values. When the confounders are missing not at random, causal effects are generally not identifiable. In this article, we propose a novel framework for nonparametric identification of causal effects with confounders missing not at random, but subject to instrumental missingness, that is, the missing data mechanism is independent of the outcome, given the treatment and possibly missing confounder values. We also give a nonparametric two-stage least squares estimator of the average causal effect based on series approximation, which overcomes an ill-posed inverse problem by restricting the estimation space to a compact subspace. The simulation studies show that our estimator can correct for confounding bias when confounders are subject to instrumental missingness.

Keywords: Completeness; Fredholm integral equation; Ill-posed inverse problem; Kernel-based estimator; Missing not at random.

1 Introduction

Observational studies are often used to infer causal effects in medical and social science studies. Under the assumption that the treatment-outcome relationship is unconfounded,

*Department of Statistics, North Carolina State University, North Carolina 27695, U.S.A.

†Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, U.S.A.

‡Department of Statistics, University of California, Berkeley, California 94720, U.S.A.

and that the confounders are fully observed, current causal inference techniques, including propensity score matching, subclassification and weighting (Imbens and Rubin; 2015), can effectively adjust for confounding in observational studies. Although commonly present in practice, much less work has been done in settings when confounders have missing values. A complete-case analysis that excludes units with missing values can be biased and inefficient.

To handle confounders with missing values, Rosenbaum and Rubin (1984) and D’Agostino Jr and Rubin (2000) developed a generalized propensity score approach. Their approach is based on a modified ignorability assumption, which states that to remove all confounding bias, it is sufficient to adjust for the missing pattern and the observed values of confounders. Under this assumption, causal effects are identifiable and the balancing property of the propensity score (Rosenbaum and Rubin; 1983) with fully observed confounders carries over to the generalized propensity score. Estimation of causal effects can then be based on classical propensity score methods. However, the modified ignorability assumption suggests that units may have different confounders depending on the missing pattern, which is often questionable in practice. Another common approach assumes that the missing data mechanism depends only on the observed data, that is, the confounders are missing at random (Rubin; 1976). Under this assumption, the full data distribution is identifiable, which facilitates likelihood and Bayesian inferences. Previous works have used multiple imputation (Rubin; 1976, 1987) for causal inference with missing covariates; see, for example, Qu and Lipkovich (2009), Crowe et al. (2010), Mitra and Reiter (2011), and Seaman and White (2014). Under the missing at random assumption, multiple imputation can generally provide reasonably good estimates (Schafer; 1997). However, in practice, the missing pattern may depend on the missing values themselves, a scenario commonly known as missing not at random (Rubin and Little; 2002). In this case, the aforementioned multiple imputation methods may fail to provide valid inference. We refer interested readers to Mattei (2009) for a comparison of complete-case analysis, propensity score matching, and multiple imputation on a real-life data example.

Causal inference with confounders missing not at random is challenging, because without further assumptions, neither causal effects nor the full data distribution are identifiable in general. Without a formal characterization of identification conditions, whether or not the subsequent statistical analyses produce meaningful results is not apparent. As far as we are

aware, there has not been much discussion on identification of causal effects in this setting. Prior to our work, Ding and Geng (2014) investigated this problem, but their approach is only applicable to the discrete case, which is restrictive both theoretically and practically. In this article, we provide a general framework for nonparametric identification of causal effects with confounders subject to instrumental missingness, that is, the missing pattern is independent of the outcome, given the treatment and possibly missing confounders. We formulate the identification problem based on an integral equation, and show, using the notion of completeness of density functions, that the full data distribution is identifiable. As a result, our framework also yields identification of the average causal effect under the assumption of no unmeasured confounding. Moreover, we develop a nonparametric two-stage least squares estimator of the average causal effect based on series approximation, which overcomes an ill-posed inverse problem by restricting the estimation space to a compact subspace.

2 Setup, notation, and assumptions

2.1 Potential outcomes, causal effects, and ignorability

Following Neyman (1923) and Rubin (1974), we use the potential outcomes framework. Suppose that the treatment is a binary variable $A \in \{0, 1\}$, with 0 and 1 being the labels for control and active treatments, respectively. For each level of treatment a , we assume that there exists a potential outcome $Y(a)$, representing the outcome had the subject, possibly contrary to the fact, been given treatment a . The observed outcome is $Y = Y(A) = AY(1) + (1 - A)Y(0)$. Let $X = (X_1, \dots, X_p)$ be a vector of p -dimensional pre-treatment covariates. We assume that a sample of size n consists of independent and identically distributed draws from the distribution of $\{A, X, Y(0), Y(1)\}$. The covariate-specific causal effect is $\tau(X) = E\{Y(1) - Y(0) \mid X\}$, and the average causal effect is $\tau = E\{\tau(X)\} = E\{Y(1) - Y(0)\}$. We focus on τ , and a similar discussion applies to the average causal effect on the treated $\tau_{\text{ATT}} = E\{Y(1) - Y(0) \mid A = 1\}$. These causal effects cannot be identified without further assumptions, because for each subject, only one potential outcome is observed. The following assumptions are commonly made in causal inference (Rosenbaum and Rubin; 1983).

Assumption 1 (Ignorability) $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X$.

Assumption 2 (Overlap) *With probability 1, there exist constants c_1 and c_2 such that $0 < c_1 \leq e(X) \leq c_2 < 1$, where $e(X) = \text{pr}(A = 1 \mid X)$ is the propensity score.*

It is well known that under Assumptions 1 and 2, adjusting for the propensity score removes confounding bias (Rosenbaum and Rubin; 1983). In practice, the causal effect τ can be estimated consistently through propensity score matching, subclassification or weighting. See Imbens and Rubin (2015) for a textbook discussion.

2.2 Confounders with missing values

In this article, we consider the case where X contains missing values. Let $R = (R_1, \dots, R_p)$ be the vector of missing indicators, that is, $R_j = 1$ if the j th component X_j is observed and 0 if it is missing. Let \mathcal{R} be the set of all possible values of R . For simplicity, we use $R = 1_p$ to denote the p -vector of 1's, and 0_p to denote the p -vector of 0's. We write $X = (X_{\text{obs}}, X_{\text{mis}})$, where X_{obs} and X_{mis} represent the observed and missing parts, respectively. In this setting, the propensity score may not be identifiable due to missing values in X . Instead, Rosenbaum and Rubin (1984) made the following modified ignorability assumption.

Assumption 3 (Observed ignorability) $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid (X_{\text{obs}}, R)$.

Under Assumption 3, $\tau = E\{E(Y \mid A = 1, X_{\text{obs}}, R)\} - E\{E(Y \mid A = 0, X_{\text{obs}}, R)\}$ is identifiable. Rosenbaum and Rubin (1984) also defined the generalized propensity score as $e(X_{\text{obs}}, R) = \text{pr}(A = 1 \mid X_{\text{obs}}, R)$, and showed that $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid e(X_{\text{obs}}, R)$. Therefore, classical propensity score methods can be applied to estimate τ . The advantage of this approach is that no assumptions on the missing data mechanism of X is necessary for identification of τ . However, this approach suffers from several drawbacks. First, under Assumption 3, a pre-treatment covariate may be a confounder when it is observed, but is not a confounder when it is missing. This is often hard to justify scientifically. Second, the observed ignorability assumption could be dubious if missingness happens after the treatment assignment, that is, R is a post-treatment variable affected by A . Third, estimation of the generalized propensity score can be challenging. It is often the case that some missing

patterns have few observations, especially if there are many covariates possibly missing. Analysts hence have to make strong parametric assumptions to leverage information across different missing data patterns (D’Agostino Jr and Rubin; 2000).

2.3 Missing data mechanisms for the confounders

Assume that the distribution of (A, X, Y, R) is absolutely continuous with respect to some measure, with $f(x, y, R = r, A = a)$ being the density or probability mass function of the appropriate variables. The following identity relates the full data distribution to the observed data distribution:

$$f(x, y, R = 1_p \mid A = a) = f(x, y \mid A = a) \text{pr}(R = 1_p \mid x, y, A = a). \quad (1)$$

In our discussion, identification of the full data distribution means that it is uniquely determined by the observed data distribution. Given (1), identification of $f(x, y \mid A = a)$ can be achieved through identification of $\text{pr}(R = 1_p \mid x, y, A = a)$. If the missing at random assumption that $R \perp\!\!\!\perp X_{\text{mis}} \mid (A, X_{\text{obs}}, Y)$ holds, then $\text{pr}(R = 1_p \mid x, y, A = a) = \text{pr}(R = 1_p \mid x_{\text{obs}}, y, A = a)$ is identifiable. However, missing at random may not be plausible if, as is likely in practice, the missing pattern is related to the missing values of confounders. Instead, we consider the following missing data mechanism.

Assumption 4 (Instrumental missingness) $R \perp\!\!\!\perp Y \mid (A, X)$.

Under Assumption 4, Y satisfies the exclusion property of an instrumental variable for R (Zhao and Shao; 2015), motivating the name instrumental missingness. This assumption is most plausible for prospective observational studies that have X measured long before the outcome takes place. Figure 1 provides a directed acyclic graph (Pearl; 2009) that implies Assumptions 1 and 4. Notably, A and Y have no common parents except for X , encoding Assumption 1; R and Y have no common parents, encoding Assumption 4. In our framework, we allow for unmeasured common causes of R and A , and the dependence of R on the missing part of X . Moreover, unlike Assumption 3, we allow for the possibility that R is a post-treatment variable affected by A .

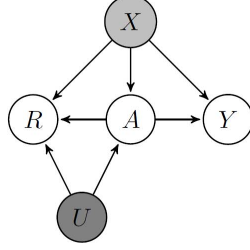


Figure 1: A direct acyclic graph illustrating the dependence of variables under Assumptions 1 and 4. White nodes represent observed variables, the light grey node represents the variable with missing values, and the dark node represents unmeasured variables.

3 Nonparametric identification

3.1 Integral equation representation

We now consider identification of the distribution of (A, X, Y, R) and the average causal effect. As discussed in §2.3, identification of $f(x, y \mid A = a)$ relies on identification of $\text{pr}(R = 1_p \mid x, y, A = a)$. Note further that

$$\text{pr}(R = r \mid x, y, A = a) = \frac{\text{pr}(R = r \mid x, y, A = a)}{\sum_{r \in \mathcal{R}} \text{pr}(R = r \mid x, y, A = a)} = \frac{\xi_{ra}(x, y)}{\sum_{r \in \mathcal{R}} \xi_{ra}(x, y)}, \quad (2)$$

where

$$\xi_{ra}(x, y) = \frac{\text{pr}(R = r \mid x, y, A = a)}{\text{pr}(R = 1_p \mid x, y, A = a)} \quad (r \in \mathcal{R}, a = 0, 1). \quad (3)$$

We then only need to identify $\xi_{ra}(x, y)$. On the other hand, $\xi_{ra}(x, y)$ connects the observed data distribution $f(x_{\text{obs}}, y, R = r \mid A = a)$ and the complete case distribution $f(x, y, R = 1_p \mid A = a)$ through the following equation:

$$\begin{aligned} f(x_{\text{obs}}, y, R = r \mid A = a) &= \int f(x, y, R = r \mid A = a) d\nu(x_{\text{mis}}) \\ &= \int \frac{\text{pr}(R = r \mid x, y, A = a)}{\text{pr}(R = 1_p \mid x, y, A = a)} f(x, y, R = 1_p \mid A = a) d\nu(x_{\text{mis}}) \\ &= \int \xi_{ra}(x, y) f(x, y, R = 1_p \mid A = a) d\nu(x_{\text{mis}}), \end{aligned} \quad (4)$$

where $\nu(x_{\text{mis}})$ is the Lebesgue measure for continuous x_{mis} and the counting measure for discrete x_{mis} . In (4), $f(x_{\text{obs}}, y, R = r \mid A = a)$ and $f(x, y, R = 1_p \mid A = a)$ are identifiable from the observed data distribution. We have thus turned identification of $\xi_{ra}(x, y)$ to the problem of solving $\xi_{ra}(x, y)$ from (4), which however requires further assumptions. We illustrate this issue in the following example.

Example 1 Suppose that X and Y are discrete, and that $X_j \in \{x_{j1}, \dots, x_{jJ_j}\}$ for $j = 1, \dots, p$ and $Y \in \{y_1, \dots, y_K\}$. Then, (4) can be written as a linear system. For $r = 0_p$ and a given a , the left hand side of (4) has K equations, whereas the right hand side of (4) has $K \times J$ unknown quantities $\xi_{ra}(x, y)$, where $J = J_1 \times \dots \times J_p$. As $K < K \times J$, $\xi_{ra}(x, y)$ is generally not identifiable for $r = 0_p$.

Under Assumption 4,

$$\xi_{ra}(x, y) = \xi_{ra}(x) = \frac{\text{pr}(R = r \mid x, A = a)}{\text{pr}(R = 1_p \mid x, A = a)}. \quad (5)$$

For Example 1, (5) reduces the number of unknown quantities on the right hand side of (4) to J . It is then possible to identify $\xi_{ra}(x)$ if Y has at least as many categories as X , that is, $K \geq J$, which intuitively means that we need more information in Y than X . To solve $\xi_{ra}(x)$ from (4), we also need the linear system to be non-degenerate, or the $K \times J$ matrix with the (x, y) th component $f(x, y, R = 1_p \mid A = a)$ to be of full column rank.

Proposition 1 Under Assumption 4, suppose that X and Y are discrete, and that $X_j \in \{x_{j1}, \dots, x_{jJ_j}\}$ for $j = 1, \dots, p$ and $Y \in \{y_1, \dots, y_K\}$. The distribution of (A, X, Y, R) is identifiable if for any a , $\text{Rank}(\Theta_a) = J$, where Θ_a is the $K \times J$ matrix with the (x, y) th component $f(x, y, R = 1_p \mid A = a)$.

For the general case where X and Y are not necessarily discrete, combining (4) and (5) yields the following integral equation, which is the basis of our identification framework.

Proposition 2 Under Assumption 4, for any r and a , the following equation

$$f(x_{\text{obs}}, y, R = r \mid A = a) = \int \xi_{ra}(x) f(x, y, R = 1_p \mid A = a) d\nu(x_{\text{mis}}) \quad (6)$$

holds for any x_{obs} and y .

3.2 Completeness and identification of the full data distribution

To generalize the rank condition in Proposition 1 that ensures unique existence of $\xi_{ra}(x)$, we use the notion of completeness, which is closely related to the concept of a complete statistic (Lehmann and Scheffé; 1950; Basu; 1955).

Definition 1 (Completeness) *A function $f(x, y)$ is complete in y , if given any squared integrable function $g(x)$, $\int g(x)f(x, y)d\nu(x) = 0$ for any y implies $g(x) = 0$ almost surely.*

The condition of completeness is not restrictive (Chen et al.; 2014). It holds for many common models, such as generalized linear models. See Blundell et al. (2007) for additional examples. For illustration, we give sufficient conditions for the completeness of distribution functions in an exponential family.

1 *The distribution $f(x, y) = \psi(x)h(y) \exp\{\lambda(y)^\top \eta(x)\}$ is complete in y if $\psi(x) > 0$, $\lambda(y) > 0$ for $y \in \mathcal{B}$ when \mathcal{B} is an open set, and the mapping $x \mapsto \eta(x)$ is one-to-one.*

The completeness of distribution functions has been used for identification in many scenarios. Newey and Powell (2003) and D'Haultfoeulle (2011) used similar conditions for identification of nonparametric instrumental variable regression models, while Carroll et al. (2010) and An and Hu (2012) specified completeness conditions for identification of measurement error models. Our article is the first to use completeness of distribution functions for causal inference with confounders missing not at random.

Assumption 5 *The joint distribution $f(x, y, R = 1_p \mid A = a)$ is complete in y , for $a = 0, 1$.*

When X and Y are discrete with finite supports, Assumption 5 is equivalent to the rank condition in Proposition 1. On the one hand, since the condition $\int g(x)f(x, y, R = 1_p \mid A = a)d\nu(x) = 0$ is made for all levels of y , there are K such conditions; on the other hand, $g(x) = 0$ has J restrictions. Hence, a necessary condition for Assumption 5 is $K \geq J$. See the Supplementary Material for a formal discussion. Assumption 5 is testable for the discrete case. For example, if y is binary, then the rank condition is equivalent to a testable condition that $Y \not\perp X \mid (R = 1_p, A = a)$. Assumption 5, however, is untestable for the continuous case with infinite-dimensional nonparametric models (Canay et al.; 2013). Thus,

while the completeness condition provides an intuitive generalization of the rank condition, the empirical implications of these conditions are substantially different in this regard.

Assumption 5 is sufficient to ensure the unique existence of $\xi_{ra}(x)$ from (6). To see this, suppose that both $\xi_{ra}^1(x)$ and $\xi_{ra}^2(x)$ satisfy (6), then $\int \{\xi_{ra}^1(x) - \xi_{ra}^2(x)\} f(x, y, R = 1_p | A = a) d\nu(x_{\text{mis}}) = 0$ for all y and x_{obs} . Integrating this equation with respect to X_{obs} , we have $\int \{\xi_{ra}^1(x) - \xi_{ra}^2(x)\} f(x, y, R = 1_p | A = a) d\nu(x) = 0$ for all y . By completeness in Assumption 5, $\xi_{ra}^1(x) = \xi_{ra}^2(x)$ almost surely.

Theorem 1 *Under Assumptions 4 and 5, the distribution of (A, X, Y, R) is identifiable.*

Remark 1 *Assumption 2 is implied by Assumption 4. We illustrate the idea with discrete X and Y , and leave the formal discussion to the Supplementary Material. Suppose that there exists x^* with $\text{pr}(X = x^*) > 0$, such that $e(x^*) = \text{pr}(A = 1 | X = x^*) = 0$. Then,*

$$f(x^*, y, R = 1_p | A = 1) = \frac{\text{pr}(A = 1 | X = x^*) f(x^*, y) \text{pr}(R = 1_p | x^*, y, A = 1)}{\text{pr}(A = 1)} = 0,$$

for any y , which indicates that one column in Θ_1 is zero. Therefore, Θ_1 is not of full column rank, violating the completeness condition.

Remark 2 *In the presence of completely unmeasured confounders, instrumental variable methods can be used to identify causal effects; see, for example, Angrist et al. (1996) and Hernán and Robins (2006). Our identification framework, however, distinguishes from the instrumental variable settings. This is because in the instrumental variable settings where there are completely unmeasured confounders, $f(x, y, R = 1_p | A = a) = 0$ for all x and y , implying that it is not complete in y .*

3.3 Nonparametric identification of the causal effect

Under Assumptions 1, 4 and 5, identification of the average causal effect τ can be achieved in two steps. First, under Assumptions 1 and 4, $\tau(X)$ can be identified by

$$\tau(X) = E(Y | X, A = 1, R = 1_p) - E(Y | X, A = 0, R = 1_p). \quad (7)$$

Second, under Assumptions 4 and 5, as shown in Theorem 1, the distribution of (A, X, Y, R) is identifiable. The marginal distribution of X , $f(x)$, is hence also identifiable. Some algebra yields

$$f(x) = \sum_{a=0}^1 f(x \mid A = a) \text{pr}(A = a) = \sum_{a=0}^1 \frac{\text{pr}(A = a, R = 1_p)}{\text{pr}(R = 1_p \mid x, A = a)} f(x \mid A = a, R = 1_p). \quad (8)$$

Theorem 2 gives the identification formula for τ .

Theorem 2 *Under Assumptions 1, 4 and 5, the average causal effect τ is identified by*

$$\tau = \int \tau(x) f(x) dx = \sum_{a=0}^1 \text{pr}(A = a, R = 1_p) \int \tau(x) \frac{f(x \mid A = a, R = 1_p)}{\text{pr}(R = 1_p \mid x, A = a)} dx, \quad (9)$$

where $\tau(x)$ is identified by (7), $\text{pr}(A = a, R = 1_p)$ and $f(x \mid A = a, R = 1_p)$ depend only on the observed data, and $\text{pr}(R = 1_p \mid x, A = a)$ can be identified by (2), (5) and (6), for $a = 0$ and 1.

4 Nonparametric estimation of the average causal effect

We consider nonparametric estimation of the average causal effect τ . According to (9), estimation of τ can be based on estimation of $\tau(x)$, $\text{pr}(A = a, R = 1_p)$, $f(x \mid A = a, R = 1_p)$ and $\text{pr}(R = 1_p \mid x, A = a)$. In the following, we use $\hat{\tau}(x)$, $\hat{\text{pr}}(A = a, R = 1_p)$, and $\hat{f}(x \mid A = a, R = 1_p)$ to denote the spline estimator of $\tau(x)$, the frequency estimator of $\text{pr}(A = a, R = 1_p)$ and the kernel density estimator of $f(x \mid A = a, R = 1_p)$, respectively. The key then is to estimate $\text{pr}(R = 1_p \mid x, A = a)$ or equivalently $\xi_{ra}(x)$ from (6). In (6), replacing $f(x_{\text{obs}}, y, R = r \mid A = a)$ and $f(x, y, R = 1_p \mid A = a)$ by the corresponding nonparametric estimators $\hat{f}(x_{\text{obs}}, y, R = r \mid A = a)$ and $\hat{f}(x, y, R = 1_p \mid A = a)$, we obtain

$$\hat{f}(x_{\text{obs}}, y, R = r \mid A = a) = \int \xi_{ra}(x) \hat{f}(x, y, R = 1_p \mid A = a) d\nu(x_{\text{mis}}), \quad (10)$$

which is a Fredholm integral equation of the first kind. There are several challenges in solving (10). First, although, as we show in §3, the population equation (6) has a unique solution,

the sample equation (10) may not. Second, $\xi_{ra}(x)$ is an infinite-dimensional parameter, estimation of which often relies on some approximation. Lastly, solving $\xi_{ra}(x)$ from (10) is an ill-conditioned problem, meaning that even a slight perturbation of $\hat{f}(x_{\text{obs}}, y, R = r \mid A = a)$ and $\hat{f}(x, y, R = 1_p \mid A = a)$ can lead to a large variation in $\xi_{ra}(x)$. As a result, plugging in consistent estimators of $f(x_{\text{obs}}, y, R = r \mid A = a)$ and $f(x, y, R = 1_p \mid A = a)$ to (6) does not necessarily yield a consistent estimator of $\xi_{ra}(x)$.

To solve the existence and computation issues, we use series approximation (Kress et al.; 1999) and least squares estimation. Suppose $\xi_{ra}(x)$ can be approximated by the Hermite polynomials:

$$\xi_{ra}(x) \approx \sum_{j=1}^J \gamma_{ra}^j h_j(\tilde{x}), \quad h_j(\tilde{x}) = \exp(-\tilde{x}^\top \tilde{x}) \tilde{x}^{\lambda_j}, \quad (11)$$

where $\tilde{x} = \Sigma_x^{-1/2}(x - \mu_x)$, and μ_x and Σ_x are the mean and covariance matrix of X , respectively, and λ_j is increasing in j . Here, the set $\{h_j(x) : j = 1, \dots, J\}$ forms a Hermite polynomial basis. Substituting the approximation of $\xi_{ra}(x)$ in (10), the vector of coefficients $\gamma_{ra} = (\gamma_{ra}^1, \dots, \gamma_{ra}^J)^\top$ can be estimated by minimizing the following residual sum of squares:

$$Q(\gamma_{ra}) = \sum_{i=1}^n \left[\hat{f}(x_{\text{obs},i}, y_i, R_i = r \mid A_i = a) - \sum_{j=1}^J \gamma_{ra}^j \hat{E}\{h_j(\hat{x}) \mid x_{\text{obs},i}, y_i, A_i = a, R_i = 1_p\} \hat{f}(x_{\text{obs},i}, y_i, R_i = 1_p \mid A_i = a) \right]^2, \quad (12)$$

where $\hat{x} = \hat{\Sigma}_x^{-1/2}(x - \hat{\mu}_x)$ is an estimate of \tilde{x} , $\hat{\mu}_x$ and $\hat{\Sigma}_x$ are the empirical mean and covariance matrix of X , respectively, and $\hat{E}\{h_j(\hat{x}) \mid x_{\text{obs}}, y, A = a, R = 1_p\}$ is a nonparametric estimator of $E\{h_j(\tilde{x}) \mid x_{\text{obs}}, y, A = a, R = 1_p\}$. For clarification, the expectation in $E\{h_j(\tilde{x}) \mid x_{\text{obs}}, y, A = a, R = 1_p\}$ is taken with respect to $f(x_{\text{mis}} \mid x_{\text{obs}}, y, A = a, R = 1_p)$, where $(x_{\text{obs}}, x_{\text{mis}})$ is determined by the missing pattern $R = r$.

To solve the noncontinuity issue, we impose regularization conditions. Previous works include Tikhonov's regularization (Honerkamp and Weese; 1990), restriction to compact spaces (Newey and Powell; 2003), among others. We restrict the parameter space of $\hat{\xi}_{ra}(x)$ to a compact subspace, which regularizes the problem to be well posed. This is because

integration is a continuous operator, and restricting to a compact subspace makes its inverse be continuous. Let Λ be some positive definite $J \times J$ matrix and B be a positive constant. We then obtain γ_{ra} by minimizing $Q(\gamma_{ra})$ in (12), subject to the constraint $\gamma_{ra}^T \Lambda \gamma_{ra} \leq B$. This constraint controls the variance of $\hat{\xi}_{ra}(x)$. The regularization details are presented in the Supplementary Material. An estimator of $\xi_{ra}(x)$ is then

$$\hat{\xi}_{ra}(x) = \sum_{j=1}^J \hat{\gamma}_{ra}^j h_j(\hat{x}). \quad (13)$$

The probability $\text{pr}(R = 1_p \mid x, A = a)$ can then be estimated by $\hat{\text{pr}}(R = 1_p \mid x, A = a) = \{1 + \sum_{r \neq 1_p} \hat{\xi}_{ra}(x)\}^{-1}$.

Finally, τ can be estimated by applying a numerical approximation technique for

$$\sum_{a=0}^1 \hat{\text{pr}}(A = a, R = 1_p) \int \hat{\tau}(x) \frac{\hat{f}(x \mid A = a, R = 1_p)}{\hat{\text{pr}}(R = 1_p \mid x, A = a)} dx. \quad (14)$$

For example, deterministic or Monte Carlo numerical integration algorithms can be used. In our simulation and data analysis, we use the importance sampling technique (Rubinstein and Kroese; 2011), as sampling directly from nonparametric density estimators is difficult.

Remark 3 (Choice of tuning parameters) *The proposed estimator depends on several tuning parameters: the number of the Hermite polynomial functions J , the bound B for regularization, and tuning parameters in the kernel-based estimators. On the one hand, J and B should be large enough to ensure that the series estimator approximates the true underlying function well; on the other hand, J and B cannot be too large, in order to control the variance of our estimator. In practice, we suggest using data-driven methods, such as cross-validation, to choose these parameters. A sensitivity analysis varying the tuning parameters is also recommended.*

The proposed estimator is consistent under regularity conditions discussed in the Supplementary Material, and its confidence interval may be constructed via the bootstrap. We use the same tuning parameters for all bootstrap samples. We relegate further technical details to the Supplementary Material.

5 Examples

5.1 Simulation studies

In the simulation study, we assess the performance of the proposed nonparametric estimator relative to existing estimators: (i) the unadjusted estimator, which simply takes difference of the average outcomes between the treated and control groups; (ii) the propensity score weighting estimator, where the propensity scores are estimated separately by logistic regressions with observed covariates for each missing pattern; (iii) multiple imputation estimators. The idea of multiple imputation is to fill the missing values m times by sampling from the posterior predictive distribution of the missing values given the observed values. Many variations have been previously proposed, with possible combinations of the following factors: (a) the imputation model using the outcome or not (Mitra and Reiter; 2011); (b) proper or improper multiple imputation (Seaman and White; 2014); (c) the propensity score model incorporating the missing pattern or not (Qu and Lipkovich; 2009). In proper multiple imputation, the analysis approach is to obtain m propensity score weighting estimates of the causal effect, and then average these estimates to get the final multiple imputation estimate. In improper multiple imputation, the analysis approach is to average each unit's m propensity scores, and then estimate the causal effect using a single set of averaged propensity scores. We consider six multiple imputation estimators specified in Table 1. Although these methods have been widely used in practice, there is little research in investigating their performance with variables missing not at random. This motivates us to compare these estimators in our simulation study using two settings: one on artificial data and one on real data.

We generate samples of size n , with $n = 400, 800$ and 1600 . The covariates are $X_i = (X_{1i}, X_{2i})$, where $X_{1i} \sim N(1, 1)$ and $X_{2i} \sim \text{Bernoulli}(0.5)$. The potential outcome variables are $Y_i(0) = 0.5 + 2X_{1i} + X_{2i} + \epsilon_{0i}$ and $Y_i(1) = 3X_{1i} + 2X_{2i} + \epsilon_{1i}$, where $\epsilon_{0i} \sim N(0, 1)$ and $\epsilon_{1i} \sim N(0, 1)$. The treatment indicator A_i is generated from $\text{Bernoulli}(\pi_i)$, where $\text{logit}(\pi_i) = 1.25 - 0.5X_{1i} - 0.5X_{2i}$. The observed outcome is $Y_i = A_iY_i(1) + (1 - A_i)Y_i(0)$. We consider that A_i , X_{2i} and Y_i are fully observed, but X_{1i} has missing values. The missing indicator of X_{1i} , R_i , is generated from $\text{Bernoulli}(p_i)$, where $\text{logit}(p_i) = -2 + 2X_{1i} + A_i(1.5 + X_{2i})$. Under

our data generating mechanism, the ignorability and instrumental missingness assumptions hold, but R depends on X_1 so that X_1 is missing not at random. The average response rate is about 67%, and the true value of average causal effect τ is 1. We compare nine estimators specified in Table 1. For our proposed estimator, $\hat{\tau}(x)$ is estimated using cubic splines with 5 knots, and the density functions are estimated by kernel-based estimators with the Gaussian kernel. The smoothing parameters in the smoothing spline estimator and the bandwidths in the kernel-based estimators are chosen by 10-fold cross-validation. For $\hat{\xi}_{ra}(x)$, the number of the Hermite polynomial basis functions is chosen to be 5, and the bound for regularization is chosen to be 50. We also vary the choices of tuning parameters; the results remain close to what we report in Table 1. See the Supplementary Material for additional simulation results. For multiple imputation estimators, the imputation size is $m = 100$. It should be noted that Rubin’s rule may not work for variance estimation, because the weighting estimator may not be self-efficient (Yang and Kim; 2016). Instead, the variance in the multiple imputation point estimate could be evaluated by the bootstrap methods (Tu and Zhou; 2002). To construct confidence intervals, we use the bootstrap with $B = 500$ samples.

Table 1 shows the simulation results over 2,000 Monte Carlo samples. The unadjusted estimator, the propensity score weighting estimator, and multiple imputation estimators are biased. As a result, the coverage rates of the confidence intervals for these methods are quite poor. Among the multiple imputation estimators, proper and improper imputations have similar performances, which is consistent with the asymptotic equivalence of the two multiple imputation estimators. Multiple imputation has the worst performance when using all observed values including the outcome for imputing the missing values of the covariate. This is because its validity relies on the missing at random assumption, which is violated in our context. In comparison, Qu and Lipkovich (2009)’s method, which includes the missing pattern in propensity score estimation, decreases biases and improves empirical coverages. This is because their method utilizes the information provided by missingness. Multiple imputation excluding the outcome variable for imputing the missing values of the covariate is better than all other alternatives of multiple imputation in our simulation context. Lastly, the proposed method has negligible biases and good coverages, and variance decreasing with the sample size.

Table 1: Simulation results: bias ($\times 10^{-2}$) and variance ($\times 10^{-3}$) of the point estimator of τ , coverage (%) of 95% confidence intervals based on 2,000 Monte Carlo samples

Method	Bias	Var	Cov	Bias	Var	Cov	Bias	Var	Cov
	$n = 400$			$n = 800$			$n = 1600$		
Unadj	-127.5	77.4	0.3	-127.4	38.0	0.0	-127.2	17.5	0.0
PSW	-55.1	42.4	22.2	-54.9	20.9	5.8	-54.4	9.5	0.4
MI1prop	41.5	35.4	40.6	41.0	15.5	9.5	40.8	7.6	0.5
MI1imp	38.3	34.6	46.6	38.0	15.0	15.2	37.5	7.4	0.9
MIMPprop	29.3	73.5	83.7	28.5	33.7	65.0	28.3	14.9	30.8
MIMPimp	26.8	72.4	86.4	26.3	33.9	71.5	26.1	15.1	40.0
MI2prop	-10.8	60.0	91.4	-9.2	28.8	91.4	-9.1	13.7	86.6
MI2imp	-13.5	55.9	90.2	-11.4	27.4	90.8	-11.1	13.1	82.5
Proposed	1.2	19.4	95.1	0.9	9.6	95.2	0.8	3.9	94.9

Unadj: the unadjusted estimator; PSW: the propensity score weighting estimator; Proposed: the proposed nonparametric estimator; For the multiple imputation estimators, MI1 and MI2 denote the imputation models to be $f(X_{\text{mis}} | A, X_{\text{obs}}, Y)$, and $f(X_{\text{mis}} | A, X_{\text{obs}})$, respectively, MIMP denotes the multiple imputation missingness pattern method of Qu and Lipkovich (2009), prop and imp denote proper and improper imputations, respectively.

5.2 The causal effect of smoking on the blood lead level

We examine a data set from the 2007–2008 U.S. National Health and Nutrition Examination Survey to estimate the causal effect of smoking on the blood lead level. The data set includes 3340 subjects consisting of 679 smokers, denoted as $A = 1$, and 2661 nonsmokers, denoted as $A = 0$. The outcome variable Y is the measured lead level in blood, with observed range from 0.18 ug/dl to 33.10 ug/dl. The covariates X include income-to-poverty level, age and gender. For details of the data set, see Hsu and Small (2013). In this data set, income-to-poverty level has missing values and other variables are completely observed. The missing rate is 6% for smokers and 9.2% for non-smokers. Missing at random is dubious, because subjects with high income-to-poverty level may be less likely to disclose their income information, so that income-to-poverty level is missing not at random. On the other hand, the income-to-poverty level is perceivably unrelated to the lead level in blood, and hence the instrumental missingness assumption is plausible. The parameter of interest is the average

causal effect τ . To compare our nonparametric estimator with the existing methods, we introduce additional missing values by generating R_i from a Bernoulli distribution with mean p_i , where $\text{logit}(p_i) = 2 - 0.6 \times \text{income}_i + 0.4A_i$. This results in 46% smokers and 29% non-smokers missing their income information. We apply the proposed procedure to obtain estimates separately for groups stratified by age and gender, and then average over the empirical distribution of age and gender.

Table 2 shows the results for the average smoking effect. We note substantial differences in the point estimates between our estimator and the competitors, which illustrates the impact of the missing data assumption can have for causal inference in the presence of missing covariates. In contrast to the existing estimators, our estimator handles the covariate missing not at random more properly. From the results, on average, smoking increases the lead level in blood by 0.51 ug/dl over the full sample.

Table 2: Point estimate, standard error by the bootstrap ($B = 1000$) and 95% confidence interval

	Est	SE	95% CI		Est	SE	95% CI
Unadj	0.83	0.10	(0.63, 1.05)	MIMPprop	0.70	0.09	(0.56, 0.90)
PSW	0.71	0.09	(0.56, 0.92)	MIMPimp	0.70	0.09	(0.56, 0.91)
MI1prop	0.72	0.09	(0.58, 0.93)	MI2prop	0.73	0.09	(0.58, 0.95)
MI1imp	0.72	0.09	(0.58, 0.94)	MI2imp	0.73	0.09	(0.58, 0.95)
Proposed	0.51	0.08	(0.36, 0.70)				

6 Discussion

In our context, Y can be viewed as a shadow variable (d’Haultfoeuille; 2010) of X in the sense that Y is associated with X but is independent of R conditional on A and X . This viewpoint provides insights for generalizing our results to the case where both X and Y are missing not at random. Let $R = (R_X, R_Y)$, where R_X and R_Y are the missing indicators for X and Y , respectively. Figure 2 illustrates a shadow variable approach to causal inference in this context. Here, Z is a shadow variable of (X, Y) in the sense that $Z \not\perp (X, Y)$ and $Z \perp R \mid (A, X, Y)$. To identify the missing probability, $\text{pr}(R_X = r, R_Y = \delta \mid x, y, A = a)$, we

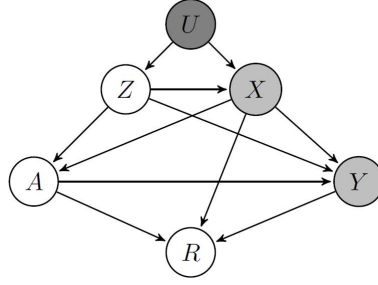


Figure 2: A direct acyclic graph illustrating a shadow variable approach for causal inference with missing confounders and outcome. White nodes represent observed variables, light grey nodes represent variables with missing values, and the dark node represents unmeasured variables.

note

$$\xi_{(r,\delta)a}(x, y) = \frac{\text{pr}(R_X = r, R_Y = \delta \mid x, y, A = a)}{\text{pr}(R_X = 1_p, R_Y = 1 \mid x, y, A = a)} \quad (r \in \mathcal{R}, \delta = 0, 1).$$

On the other hand, $\xi_{(r,\delta)a}(x, y)$ connects the observed data distribution $f(x_{\text{obs}}, z, R_X = r, R_Y = \delta \mid A = a)$ and the complete case distribution $f(x, y, z, R = 1_{p+1} \mid A = a)$ through the following equations:

$$\begin{aligned} f(x_{\text{obs}}, z, R_X = r, R_Y = 0 \mid A = a) &= \int \int \xi_{(r,0)a}(x, y) f(x, y, z, R = 1_{p+1} \mid A = a) d\nu(x_{\text{mis}}) dy, \\ f(x_{\text{obs}}, z, R_X = r, R_Y = 1 \mid A = a) &= \int \xi_{(r,1)a}(x, y) f(x, y, z, R = 1_{p+1} \mid A = a) d\nu(x_{\text{mis}}). \end{aligned}$$

One can then impose completeness on $f(x, y, z, R = 1_{p+1} \mid A = a)$ in z to guarantee the unique existence of $\xi_{(r,\delta)a}(x, y)$. Following a similar derivation in §3, the distribution of (A, X, Y, Z, R) and the causal effects can be identified.

Like many other nonparametric estimators, the proposed estimator suffers from the curse of dimensionality. Parametric or semiparametric methods can be used to overcome the problem. There exist various constructions of doubly robust estimators for the average causal effect when confounders are fully observed; see Rotnitzky and Vansteelandt (2015) for a review. However, doubly robust estimators with confounders missing not at random have not been studied in the literature. In a non-causal context, Miao and Tchetgen Tchetgen

(2016) proposed semiparametric estimators, including doubly robust estimators, of the mean of Y under missingness not at random with a shadow variable. It remains an interesting avenue for future research to investigate doubly robust estimators of the average causal effect τ .

Acknowledgment

The authors thank Eric Tchetgen Tchetgen for valuable discussions.

Supplementary material

Supplementary material includes proofs, as well as further discussions on constructions and asymptotic properties of the nonparametric estimator.

References

- An, Y. and Hu, Y. (2012). Well-posedness of measurement error models for self-reported data, *J. Econometrics* **168**: 259–269.
- Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables, *J. Am. Stat. Assoc.* **91**: 444–455.
- Basu, D. (1955). On statistics independent of a complete sufficient statistic, *Sankhyā* **15**: 377–380.
- Blundell, R., Chen, X. and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves, *Econometrica* **75**: 1613–1669.
- Canay, I. A., Santos, A. and Shaikh, A. M. (2013). On the testability of identification in some nonparametric models with endogeneity, *Econometrica* **81**: 2535–2559.

- Carroll, R. J., Chen, X. and Hu, Y. (2010). Identification and estimation of nonlinear models using two samples with nonclassical measurement errors, *J. Nonparametr. Stat.* **22**: 379–399.
- Chen, X., Chernozhukov, V., Lee, S. and Newey, W. K. (2014). Local identification of nonparametric and semiparametric models, *Econometrica* **82**: 785–809.
- Crowe, B. J., Lipkovich, I. A. and Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome, *Pharm. Stat.* **9**: 269–279.
- D’Agostino Jr, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data, *J. Am. Stat. Assoc.* **95**: 749–759.
- Deheuvels, P. (2000). Uniform limit laws for kernel density estimators on possibly unbounded intervals, *Recent Advances in Reliability Theory: Methodology, Practice and Inference*, Springer, Birkhauser, Basel, pp. 477–492.
- d’Haultfoeuille, X. (2010). A new instrumental method for dealing with endogenous selection, *J. Econometrics* **154**: 1–15.
- D’Haultfoeuille, X. (2011). On the completeness condition in nonparametric instrumental problems, *Econometric Theory* **27**: 460–471.
- Ding, P. and Geng, Z. (2014). Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates, *Stat. Med.* **33**: 1121–1133.
- Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation, *Econometrica* **55**: 363–390.
- Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators, *Annales de l’IHP Probabilités et Statistiques*, Vol. 38, pp. 907–921.
- Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: an epidemiologist’s dream?, *Epidemiology* **17**: 360–372.

- Honerkamp, J. and Weese, J. (1990). Tikhonovs regularization method for ill-posed problems, *Continuum Mech. Thermodyn.* **2**: 17–30.
- Hsu, J. Y. and Small, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies, *Biometrics* **69**: 803–811.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press, Cambridge UK.
- Kress, R., Maz’ya, V. and Kozlov, V. (1999). *Linear Integral Equations*, 2 edn, Springer: New York.
- Lehmann, E. L. and Scheffé, H. (1950). Completeness, similar regions, and unbiased estimation: Part I, *Sankhyā* **10**: 305–340.
- Mattei, A. (2009). Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing, *Statistical Methods and Applications* **18**: 257–273.
- Miao, W. and Tchetgen Tchetgen, E. J. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable, *Biometrika* **2**: 475–482.
- Mitra, R. and Reiter, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models, *Stat. Med.* **30**: 627–641.
- Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators, *J. Econometrics* **79**: 147–168.
- Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models, *Econometrica* **71**: 1565–1578.
- Neyman, J. (1923). Sur les applications de la thar des probabilités aux expériences Agricales: Essay de principe. English translation of excerpts by Dabrowska, D. and Speed, T., *Statist. Sci.* **5**: 465–472.
- Pearl, J. (2009). *Causality*, 2 edn, Cambridge: Cambridge University Press.

- Qu, Y. and Lipkovich, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach, *Stat. Med.* **28**: 1402–1414.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**: 41–55.
- Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score, *J. Am. Stat. Assoc.* **79**: 516–524.
- Rotnitzky, A. and Vansteelandt, S. (2015). Double-robust methods, in A. Tsiatis and G. Verbeke (eds), *Handbook of Missing Data Methodology*, Boca Raton, FL: CRC Press., pp. 185–212.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies., *J. Educ. Psychol.* **66**: 688–701.
- Rubin, D. B. (1976). Inference and missing data, *Biometrika* **63**: 581–592.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.
- Rubin, D. B. and Little, R. J. (2002). *Statistical Analysis with Missing Data*, 2 edn, Hoboken: Wiley.
- Rubinstein, R. Y. and Kroese, D. P. (2011). *Simulation and the Monte Carlo Method*, New York: Wiley.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.
- Seaman, S. and White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness, *Comm. Statist. Theory Methods* **43**: 3499–3515.
- Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method, *Ann. Statist.* **12**: 898–916.
- Tu, W. and Zhou, X.-H. (2002). A bootstrap confidence interval procedure for the treatment effect using propensity score subclassification, *Health Services and Outcomes Research Methodology* **3**: 135–147.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer.

Yang, S. and Kim, J. K. (2016). A note on multiple imputation for method of moments estimation, *Biometrika* **103**: 244–251.

Zhao, J. and Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data, *J. Am. Stat. Assoc.* **110**: 1577–1590.

Supplementary material

§S7 includes proofs of Proposition 1, the equivalence of completeness and the rank condition for discrete variables, Lemma 1 and Remark 1, §S8 includes the regularization of the series estimator and the asymptotic properties of the nonparametric estimators, and §S9 includes additional simulations.

S7 Proofs

S7.1 Proof of Proposition 1

We prove the result for $p = 2$. Proofs for other values of p are similar and hence omitted. For discrete covariates with $R = (0, 0)$, (5) reduces to

$$f\{y, R = (0, 0) \mid A = a\} = \sum_{i=1}^{J_1} \sum_{j=1}^{J_2} \frac{\text{pr}\{R = (0, 0) \mid x_{1i}, x_{2j}, A = a\}}{\text{pr}\{R = (1, 1) \mid x_{1i}, x_{2j}, A = a\}} \times f\{x_{1i}, x_{2j}, y, R = (1, 1) \mid A = a\}, \quad (\text{S1})$$

for all y and a . In a matrix form, (S1) becomes

$$\begin{pmatrix} f\{y_1, R = (0, 0) \mid A = a\} \\ \vdots \\ f\{y_K, R = (0, 0) \mid A = a\} \end{pmatrix}_{K \times 1} = \Theta_a \begin{pmatrix} \xi_{(0,0)a}(x_{11}, x_{21}) \\ \vdots \\ \xi_{(0,0)a}(x_{1J_1}, x_{2J_2}) \end{pmatrix}_{(J_1 J_2) \times 1}, \quad (\text{S2})$$

where

$$\Theta_a = \begin{pmatrix} f\{x_{11}, x_{21}, y_1, R = (1, 1) \mid A = a\} & \cdots & f\{x_{1J_1}, x_{2J_2}, y_1, R = (1, 1) \mid A = a\} \\ \vdots & \ddots & \vdots \\ f\{x_{11}, x_{21}, y_K, R = (1, 1) \mid A = a\} & \cdots & f\{x_{1J_1}, x_{2J_2}, y_K, R = (1, 1) \mid A = a\} \end{pmatrix}_{K \times (J_1 J_2)},$$

and

$$\xi_{(0,0)a}(x_1, x_2) = \frac{\text{pr}\{R = (0, 0) \mid x_{1i}, x_{2j}, A = a\}}{\text{pr}\{R = (1, 1) \mid x_{1i}, x_{2j}, A = a\}}.$$

In the linear system (S2), the vector on the left hand side and the coefficients in Θ_a on the right hand side depend only on the observed data, and therefore are identifiable. The linear system for the $\xi_{(0,0)a}(x_1, x_2)$'s has a unique solution if and only if Θ_a has a full column rank $J_1 J_2$. Similarly, for $R = (1, 0)$, for any x_1 and a ,

$$\begin{pmatrix} f\{x_1, y_1, R = (1, 0) \mid A = a\} \\ \vdots \\ f\{x_1, y_K, R = (1, 0) \mid A = a\} \end{pmatrix}_{K \times 1} = \Theta_{x_1 a} \begin{pmatrix} \xi_{(1,0)a}(x_1, x_{21}) \\ \vdots \\ \xi_{(1,0)a}(x_1, x_{2J_2}) \end{pmatrix}_{J_2 \times 1}, \quad (\text{S3})$$

where

$$\Theta_{x_1 a} = \begin{pmatrix} f\{x_1, x_{21}, y_1, R = (1, 1) \mid A = a\} & \cdots & f\{x_1, x_{2J_2}, y_1, R = (1, 1) \mid A = a\} \\ \vdots & \ddots & \vdots \\ f\{x_1, x_{21}, y_K, R = (1, 1) \mid A = a\} & \cdots & f\{x_1, x_{2J_2}, y_K, R = (1, 1) \mid A = a\} \end{pmatrix}_{K \times J_2}.$$

The linear system (S3) has a unique solution for the $\xi_{(1,0)a}(x_1, x_2)$'s if and only if for any x_1 , $\Theta_{x_1 a}$ has a column rank J_2 , which is guaranteed if Θ_a has a full column rank $J_1 J_2$. For $R = (0, 1)$, for any x_2 and a ,

$$\begin{pmatrix} f\{x_2, y_1, R = (0, 1) \mid A = a\} \\ \vdots \\ f\{x_2, y_K, R = (0, 1) \mid A = a\} \end{pmatrix}_{K \times 1} = \Theta_{x_2 a} \begin{pmatrix} \xi_{(0,1)a}(x_{11}, x_2) \\ \vdots \\ \xi_{(0,1)a}(x_{1J_1}, x_2) \end{pmatrix}_{J_1 \times 1}, \quad (\text{S4})$$

where

$$\Theta_{x_2 a} = \begin{pmatrix} f\{x_{11}, x_2, y_1, R = (1, 1) \mid A = a\} & \cdots & f\{x_{1J_1}, x_2, y_1, R = (1, 1) \mid A = a\} \\ \vdots & \ddots & \vdots \\ f\{x_{11}, x_2, y_K, R = (1, 1) \mid A = a\} & \cdots & f\{x_{1J_1}, x_2, y_K, R = (1, 1) \mid A = a\} \end{pmatrix}_{K \times J_1}.$$

The linear system (S4) has a unique solution for the $\xi_{(0,1)a}(x_1, x_2)$'s if and only if for any x_2 , $\Theta_{x_2 a}$ has a column rank J_1 , which is guaranteed if Θ_a has a full column rank $J_1 J_2$. Therefore, $\xi_{ra}(x_1, x_2)$ is identifiable if and only if Θ_a has a full column rank $J_1 J_2$.

It follows that

$$\text{pr}(R = r \mid x_1, x_2, A = a) = C_a(x_1, x_2) \xi_{ra}(x_1, x_2)$$

is identifiable, where

$$C_a(x_1, x_2) = \left\{ \sum_{r \in \mathcal{R}} \xi_{ra}(x_1, x_2) \right\}^{-1}.$$

It then follows that

$$f(x, y \mid A = a) = \frac{f(x, y, R = 1_p \mid A = a)}{\text{pr}(R = 1_p \mid x_1, x_2, A = a)}$$

is identifiable. Therefore, the joint distribution of (A, X, Y, R) , $\text{pr}(A = a)f(x, y \mid A = a)\text{pr}(R = r \mid x, A = a)$, is identifiable. This completes the proof.

S7.2 Proof of the equivalence of completeness and the rank condition for discrete variables

Suppose that $\int g(x)f(x, y, R = 1_p \mid A = a)d\nu(x) = 0$ for all $y = y_1, \dots, y_K$. For discrete X , the integral equation (5) reduces to

$$\Theta_a \begin{pmatrix} g(x_{11}, \dots, x_{p1}) \\ \vdots \\ g(x_{1J_1}, \dots, x_{pJ_p}) \end{pmatrix}_{(J_1 \times \dots \times J_p) \times 1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}_{K \times 1}. \quad (\text{S5})$$

If Θ_a is of full column rank, then the solution to the linear system (S5) is zero, that is, $g(x) = 0$ for all x , which indicates that $f(x, y, R = 1_p \mid A = a)$ is complete in y . On the other hand, if $f(x, y, R = 1_p \mid A = a)$ is complete in y , suppose that $\int g(x)f(x, y, R = 1_p \mid A = a)d\nu(x) = 0$ for all $y = y_1, \dots, y_K$, then $g(x) = 0$ for all x . In this case, the only solution to (S5) is

$$\begin{pmatrix} g(x_{11}, \dots, x_{p1}) \\ \vdots \\ g(x_{1J_1}, \dots, x_{pJ_p}) \end{pmatrix}_{(J_1 \times \dots \times J_p) \times 1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}_{(J_1 \times \dots \times J_p) \times 1}.$$

Therefore, Θ_a is of full column rank. This completes the proof.

S7.3 Proof of Lemma 1

Suppose that $\int g(x)f(x, y)dx = 0$ for all y . A bit of algebra simplifies this to

$$h(y) \int \tilde{g}(x) \exp\{\lambda(y)^\top \eta(x)\} dx = 0 \quad (\text{S6})$$

for all y , where $\tilde{g}(x) = g(x)\psi(x)$. Since the mapping $x \mapsto \eta(x)$ is one-to-one, let $t = \eta(x)$ and therefore $x = \eta^{-1}(t)$. Then, the integral equation (S6) changes to

$$h(y) \int \tilde{g}\{\eta^{-1}(t)\} [\dot{\eta}\{\eta^{-1}(t)\}]^{-1} \exp\{\lambda(y)^\top t\} dt = 0 \quad (\text{S7})$$

for all y , and particularly for all $y \in \mathcal{B}$, where $[\dot{\eta}\{\eta^{-1}(t)\}]^{-1}$ is the Jacobian matrix with $\dot{\eta}(x) = \partial\eta(x)/\partial x$. The left hand side of the integral equation (S7) as a function of $\lambda(y)$ is a multivariate Laplace transform of $\tilde{g}\{\eta^{-1}(t)\}[\dot{\eta}\{\eta^{-1}(t)\}]^{-1}$, and it cannot be zero unless $\tilde{g}\{\eta^{-1}(t)\}[\dot{\eta}\{\eta^{-1}(t)\}]^{-1}$ is zero almost everywhere. Since $[\dot{\eta}\{\eta^{-1}(t)\}]^{-1}$ is not zero, (S7) holds only if $\tilde{g}(x)$ is zero almost everywhere. Moreover, since $\psi(x)$ is not zero, $g(x)$ is zero almost everywhere. This completes the proof.

Example S2 *The Gaussian model*

$$f(x, y) = f(y | x)f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(y - \beta_0 - \beta_1^T x)^2}{2\sigma^2}\right\} f(x), \quad (\text{S8})$$

where $\beta_1 = (\beta_{11}, \dots, \beta_{1p})^T$ and $x = (x_1, \dots, x_p)^T$, is complete in y .

Using the above notation, (S8) can be expressed as $f(x, y) = \psi(x) \exp\{\lambda(y)^T \eta(x)\}$ with $\psi(x) = (2\pi\sigma^2)^{-1/2} f(x)$, $\lambda(y) = \sigma^{-2}(\beta_{11}y, \dots, \beta_{1p}y)^T$ and $\eta(x) = (x_1, \dots, x_p)^T$. Therefore, (S8) satisfies the conditions for $\lambda(y)$ and $\eta(x)$, and it is complete in y .

S7.4 Proof of Remark 1

For the continuous case, suppose that there exists a subset \mathcal{X}^* with $\text{pr}(x^* \in \mathcal{X}^*) > 0$, such that $e(x^*) = \text{pr}(A = 1 | x^*) = 0$ for any $x^* \in \mathcal{X}^*$. Following the same derivation as for the discrete case, we have, for any $x^* \in \mathcal{X}^*$ and y , that $f(x^*, y, R = 1_p | A = 1) = 0$. Then, $f(x, y, R = 1_p | A = 1)$ is not complete in y . To see this, suppose $\int g(x)f(x, y, R = 1_p | A = 1)d\nu(x) = 0$ for any y , we can let $g(x)$ be zero outside of \mathcal{X}^* but non-zero inside of \mathcal{X}^* , violating the completeness condition.

S8 Nonparametric estimation of the average causal effect

S8.1 Regularization of series estimators

Following Newey and Powell (2003), we restrict $\xi_{ra}(x)$ and its estimator $\hat{\xi}_{ra}(x)$ to belong to a compact subspace. Since the inverse of integration restricted to a compact space is continuous, this regularization turns the problem to be well-posed.

To describe the compact subspace, we first introduce additional notations. Let $\lambda = (\lambda_1, \dots, \lambda_p)^\top$ be a p -vector with non-negative integers as components, where p is the dimension of X . For any function $g(x)$, denote

$$D^\lambda g(x) = \frac{\partial^{\lambda_1}}{\partial x_1^{\lambda_1}} \cdots \frac{\partial^{\lambda_p}}{\partial x_p^{\lambda_p}} g(x),$$

and $|\lambda| = \sum_{l=1}^p \lambda_l$ gives the order of the derivative. In particular, the zero order derivative is the function itself: $D^0 g(x) = g(x)$.

For $m > 0$, $m_0, \delta_0 > p/2$, and $p/2 < \delta < \delta_0$, consider the following functional space

$$\mathcal{G}_{m,m_0,\delta_0,B} = \left\{ g(x) : \sum_{|\lambda| \leq m+m_0} \int \{D^\lambda g(\tilde{x})\}^2 (1 + \tilde{x}^\top \tilde{x})^{\delta_0} dx \leq B \right\}, \quad (\text{S9})$$

where $\tilde{x} = \Sigma_x^{-1/2}(x - \mu_x)$, μ_x and Σ_x are the mean and covariance matrix of X , respectively. Consider the norm

$$\|g\| = \max_{|\lambda| \leq m} \sup_x |D^\lambda g(\tilde{x})| (1 + \tilde{x}^\top \tilde{x})^\delta.$$

Gallant and Nychka (1987) showed that the closure of $\mathcal{G}_{m,m_0,\delta_0,B}$ with respect to the norm $\|g\|$ is compact.

Assumption S6 (Regularization of the parameter space) Assume that $\xi_{ra}(x)$ and its estimator $\hat{\xi}_{ra}(x)$ belong to $\mathcal{G}_{m,m_0,\delta_0,B}$ in (S9), for any r and a .

Remark S4 The regularization is not restrictive for the following reasons. First, by the definition of $\mathcal{G}_{m,m_0,\delta_0,B}$, the bound B requires the functions of $\mathcal{G}_{m,m_0,\delta_0,B}$ to be smooth to a certain degree and the tails of these functions to be small. In most applications, we would expect that the functions $\xi_{ra}(x)$ to be smooth and mainly concerned with the functional forms of $\xi_{ra}(x)$ over some compact region that is large enough to cover the region where observations are measured.

Given the Hermite approximation of $\xi_{ra}(x)$, the regularization in Assumption S6 becomes

$$\gamma_{ra}^T \left[\sum_{|\lambda| \leq m+m_0} \int \{D^\lambda H(\tilde{x})\} \{D^\lambda H(\tilde{x})\}^T (1 + \tilde{x}^T \tilde{x})^{\delta_0} dx \right] \gamma_{ra} \leq B, \quad (\text{S10})$$

where $\gamma_{ra} = (\gamma_{ra}^1, \dots, \gamma_{ra}^J)^T$ and $D^\lambda H(x) = (D^\lambda h_1(x), \dots, D^\lambda h_J(x))^T$. Therefore, we choose the positive definite matrix Λ in the constraint for regularization in §4 to be

$$\Lambda = \sum_{|\lambda| \leq m+m_0} \int \{D^\lambda H(\hat{x})\} \{D^\lambda H(\hat{x})\}^T (1 + \hat{x}^T \hat{x})^{\delta_0} dx,$$

where $\hat{x} = \hat{\Sigma}_x^{-1/2}(x - \hat{\mu}_x)$ is an estimate of \tilde{x} , and $\hat{\mu}_x$ and $\hat{\Sigma}_x$ are the empirical mean and covariance matrix of X , respectively. The proposed estimator $\hat{\xi}_{ra}(x)$ is obtained as in (13), where $\hat{\gamma}_{ra}$ is obtained by minimizing $Q(\gamma_{ra})$ in (12), subject to the constraint $\gamma_{ra}^T \Lambda \gamma_{ra} \leq B$.

S8.2 Computational algorithm

Our estimator for τ is summarized in the following algorithm.

Step S1 Obtain nonparametric estimators of $\tau(x)$, $f(x \mid A = a, R = 1_p)$, $f(x_{\text{obs}}, y \mid A = a, R = r)$, for all r and a . Specifically, we use

$$\hat{\tau}(x) = \hat{E}(y \mid x, A = 1, R = 1_p) - \hat{E}(y \mid x, A = 0, R = 1_p), \quad (\text{S11})$$

where $\hat{E}(y \mid x, A = a, R = 1_p)$ is a smoothing spline estimator of $E(y \mid x, A = a, R = 1_p)$, for $a = 0, 1$. Also let $\hat{f}(x \mid A = a, R = 1_p)$ and $\hat{f}(x_{\text{obs}}, y \mid A = a, R = r)$ be the kernel density estimators of $f(x \mid A = a, R = 1_p)$ and $f(x_{\text{obs}}, y \mid A = a, R = r)$, respectively.

Step S2 Obtain an series estimator of $\xi_{ra}(x)$ using the Hermite polynomials, $\hat{\xi}_{ra}(x) \approx \sum_{j=1}^J \hat{\gamma}_{ra}^j h_j(\hat{x})$. The estimator $\hat{\gamma}_{ra} = (\hat{\gamma}_{ra}^1, \dots, \hat{\gamma}_{ra}^J)^T$ is obtained by minimizing the objective function (12), subject to the constraint $\gamma_{ra}^T \Lambda \gamma_{ra} \leq B$.

Step S3 The probabilities $\text{pr}(R = 1_p \mid x, A = a)$ can be estimated by $\hat{\text{pr}}(R = 1_p \mid x, A = a) = \{1 + \sum_{r \neq 1_p} \hat{\xi}_{ra}(x)\}^{-1}$.

Step S4 The estimator of τ is obtained by (14) using a numerical approximation.

For illustration, we provide an example with a scalar X , which is subject to instrumental missingness. In this case, $R \in \mathcal{R} = \{0, 1\}$.

Example S3 In Step S1, obtain a nonparametric estimator of $\tau(x)$ as

$$\hat{\tau}(x) = \hat{E}(y \mid x, A = 1, R = 1) - \hat{E}(y \mid x, A = 0, R = 1),$$

where $\hat{E}(y \mid x, A = a, R = 1)$ is a smoothing spline estimator of $E(y \mid x, A = a, R = 1)$, for $a = 0, 1$. Also let $\hat{f}(x \mid A = a, R = 1)$ and $\hat{f}(y \mid A = a, R = 0)$ be the kernel density estimators of $f(x \mid A = a, R = 1)$ and $f(y \mid A = a, R = 0)$, respectively.

In Step S2, (10) becomes

$$\hat{f}(y, R = 0 \mid A = a) = \int \xi_{0a}(x) \hat{f}(x \mid y, A = a, R = 1) d\nu(x) \hat{f}(y, R = 1 \mid A = a),$$

where $\hat{f}(y, R = 0 \mid A = a)$ and $\hat{f}(y, R = 1 \mid A = a)$ are the kernel density estimators of $f(y, R = 0 \mid A = a)$ and $f(y, R = 1 \mid A = a)$, respectively. Obtain an series estimator of $\xi_{0a}(x)$ using the Hermite polynomials, $\hat{\xi}_{0a}(x) \approx \sum_{j=1}^J \hat{\gamma}_{0a}^j h_j(\hat{x})$. The estimator $\hat{\gamma}_{0a} = (\hat{\gamma}_{0a}^1, \dots, \hat{\gamma}_{0a}^J)^\top$ is obtained by minimizing the objective function

$$Q(\gamma_{0a}) = \sum_{i=1}^n \left[\hat{f}(y_i, R_i = 0 \mid A_i = a) - \sum_{j=1}^J \gamma_{0a}^j \hat{E}\{h_j(\hat{x}) \mid y_i, A_i = a, R_i = 1\} \hat{f}(y_i, R_i = 1 \mid A_i = a) \right]^2, \quad (\text{S12})$$

subject to the constraint $\gamma_{0a}^\top \Lambda \gamma_{0a} \leq B$, where $\hat{E}\{h_j(\hat{x}) \mid y, A = a, R = 1\}$ is a nonparametric estimator of $E\{h_j(\tilde{x}) \mid y, A = a, R = 1\}$.

In Step S3, the probabilities $\text{pr}(R = 1 \mid x, A = a)$ can be estimated by $\hat{\text{pr}}(R = 1 \mid x, A = a) = \{1 + \hat{\xi}_{0a}(x)\}^{-1}$.

In Step S4, the estimator of τ is obtained by

$$\sum_{a=0}^1 \widehat{\text{pr}}(A = a, R = 1) \int \hat{\tau}(x) \frac{\hat{f}(x \mid A = a, R = 1)}{\widehat{\text{pr}}(R = 1 \mid x, A = a)} dx, \quad (\text{S13})$$

using a numerical approximation.

S8.3 Asymptotic results

We study the consistency of the nonparametric estimators in Step S1, the series estimator of $\xi_{ra}(x)$ in Step S2, and finally the proposed estimator of τ in Step S4.

Firstly, to study the consistency of the nonparametric estimators in Step S1, we assume that the kernel functions satisfy the following regularity conditions:

Condition S1 (i) $\int_{\mathcal{R}^p} K(s) ds = 1$; (ii) $\|K\|_\infty = \sup_{x \in \mathcal{R}^p} |K(x)| = \kappa < \infty$; (iii) $K(\cdot)$ is right continuous; (iv) $\int_{\mathcal{R}^p} \Psi_K(x) dx < \infty$, where $\Psi_K(x) = \sup_{|y| \geq |x|} |K(y)|$, for $x \in \mathcal{R}^p$; and (v) the kernel function is regular and satisfies the following uniform entropy condition. Let \mathcal{K} be the class of functions indexed by x ,

$$\mathcal{K} = \left\{ K \left(\frac{x - \cdot}{h^{1/p}} \right) : h > 0, x \in \mathcal{R}^p \right\}.$$

Suppose \mathcal{B} is a Borel set in \mathcal{R}^p , and Q is some probability measure on $(\mathcal{R}^p, \mathcal{B})$. Define d_Q to be the $L_2(Q)$ -metric and $N(\epsilon, \mathcal{K}, d_Q)$ the minimal number of balls $\{g : d_Q(g, g') < \epsilon\}$ of d_Q -radius ϵ needed to cover \mathcal{K} . Let $N(\epsilon, \mathcal{K}) = \sup_Q N(\epsilon, \mathcal{K}, d_Q)$, where the supremum is taken over all probability measures Q . For some $C > 0$ and $\nu > 0$, $N(\epsilon, \mathcal{K}) \leq C\epsilon^{-\nu}$, for any $0 < \epsilon < 1$.

Sufficient conditions for Condition S1 (v) can be found in van der Vaart and Wellner (1996).

Theorem S3 (Consistency of kernel density estimators) Let $\hat{f}(x \mid A = a, R = 1_p)$ be the kernel density estimator of $f(x \mid A = a, R = 1_p)$, where the kernel function satisfies Condition S1, and the bandwidth h_n satisfies the conditions that h_n decreases to zero, h_n/h_{2n}

is bounded, $\log(1/h_n)/\log \log n \rightarrow \infty$ and $nh_n/\log n \rightarrow \infty$, as $n \rightarrow \infty$. Suppose that the true density function $f(x \mid A = a, R = 1_p)$ is bounded and uniformly continuous in x , then

$$\lim_{n \rightarrow \infty} \left\| \hat{f}(x \mid A = a, R = 1_p) - f(x \mid A = a, R = 1_p) \right\|_{\infty} = 0 \quad (\text{S14})$$

almost surely.

Theorem S4 (Consistency of kernel-based estimators for conditional means) Let $\hat{E}(y \mid x, A = a, R = 1_p)$ be the Nadaraya–Watson estimator of $E(y \mid x, A = a, R = 1_p)$, that is,

$$\hat{E}(y \mid x, A = 1, R = 1_p) = \frac{\sum_{i: R_i=1_p} A_i Y_i K\left(\frac{x-X_i}{h_n^{1/p}}\right)}{\sum_{i: R_i=1_p} A_i K\left(\frac{x-X_i}{h_n^{1/p}}\right)},$$

and

$$\hat{E}(y \mid x, A = 0, R = 1_p) = \frac{\sum_{i: R_i=1_p} (1 - A_i) Y_i K\left(\frac{x-X_i}{h_n^{1/p}}\right)}{\sum_{i: R_i=1_p} (1 - A_i) K\left(\frac{x-X_i}{h_n^{1/p}}\right)},$$

where the kernel function $K(\cdot)$ satisfies Condition S1 with support contained in $[-1/2, 1/2]^p$, and the bandwidth h_n satisfies the conditions that h_n decreases to zero, h_n/h_{2n} is bounded, $\log(1/h_n)/\log \log n \rightarrow \infty$ and $nh_n/\log n \rightarrow \infty$, as $n \rightarrow \infty$.

Let I be a compact subset of \mathcal{R}^p . For any function $\psi : \mathcal{R}^p \rightarrow \mathcal{R}$, define

$$\|\psi\|_I = \sup_{x \in I} |\psi(x)|. \quad (\text{S15})$$

Also, denote $I^\epsilon = \{x \in \mathcal{R}^p : \max_{1 \leq i \leq p} |x_i| \leq \epsilon\}$. Suppose that there exists an $\epsilon > 0$ such that $f(x \mid A = a, R = 1_p) = \int_{-\infty}^{\infty} f(y, x \mid A = a, R = 1_p) dy$ is continuous and strictly positive on I^ϵ , and that $f(y, x \mid A = a, R = 1_p)$ is continuous in x for almost every $y \in \mathcal{R}$. Suppose further that there exists an $M > 0$ such that for $X \in I^\epsilon$, $|Y| \leq M$ almost surely. Then, for any a ,

$$\lim_{n \rightarrow \infty} \left\| \hat{E}(y \mid x, A = a, R = 1_p) - E(y \mid x, A = a, R = 1_p) \right\|_I = 0 \quad (\text{S16})$$

almost surely.

A large literature has developed consistency of kernel-based estimators. The proofs of

Theorems S3 and S4 are similar to those given by Deheuvels (2000) and Giné and Guillou (2002), and therefore are omitted. The smoothing spline estimator is asymptotically equivalent to a kernel-based estimator that employs the so-called spline kernel (Silverman; 1984). It has been shown that spline kernels and Gaussian kernels satisfy Condition S1 (van der Vaart and Wellner; 1996). Therefore, by Theorems S3 and S4, the nonparametric estimators in Step S1 are consistent.

Secondly, to study the consistency of the series estimator of $\xi_{ra}(x)$ in Step S2, define the residual

$$\rho_{ra}(x, y, \gamma_{ra}) = f(x_{\text{obs}}, y, R = r \mid A = a) - \left\{ \sum_{j=1}^J \gamma_{ra}^j h_j(x) \right\} f(x, y, R = 1_p \mid A = a),$$

for any r and a , and assume the following regularity condition:

Condition S2 *$E\{||\rho_{ra}(x, y, \gamma_{ra})||^2 \mid x_{\text{obs}}, y, A = a\}$ is bounded, and there exist $M_{ra}(x, y)$ and $\nu > 0$ such that*

$$||\rho(x, y, \gamma_{ra}) - \rho(x, y, \tilde{\gamma}_{ra})|| \leq M_{ra}(x, y) ||\gamma_{ra} - \tilde{\gamma}_{ra}||^\nu,$$

for all γ_{ra} and $\tilde{\gamma}_{ra}$, and $E\{M_{ra}(x, y)^2 \mid x_{\text{obs}}, y, A = a\}$ is bounded.

Theorem S5 (Consistency of $\hat{\xi}_{ra}$) *Under Assumptions 1, 4, 5, and Condition S2, the series estimator $\hat{\xi}_{ra}(x) = \sum_{j=1}^J \hat{\gamma}_{ra}^j h_j(\hat{x})$ in (13) is consistent for $\xi_{ra}(x)$ in the sense that for $J \rightarrow \infty$ and any δ such that $p/2 < \delta < \delta_0$,*

$$\max_{|\lambda| \leq m} \sup_x \left| D^\lambda \sum_{j=1}^J \hat{\gamma}_{ra}^j h_j(\hat{x}) - D^\lambda \xi_{ra}(x) \right| (1 + x^\top x)^\delta = o_p(1).$$

Newey and Powell (2003) provided a proof for Theorem S5 in the context of instrumental variable estimation of nonparametric models. Our proof for Theorem S5 is similar, and therefore is omitted.

Finally, the consistency of the proposed estimator of the causal effect τ in Step S4 follows naturally from Theorems S3–S5.

Theorem S6 (Consistency of $\hat{\tau}$) Suppose that the conditions in Theorems S3–S5 hold. Suppose further that for some $B > 0$, $\hat{\tau}(x)$ and $\tau(x)$ are uniformly bounded for $x \in I_B = \{x : \|x\| > B\}$, and that

$$\int_{I_K^c} \frac{f(x \mid A = a, R = 1_p)}{\text{pr}(R = 1_p \mid x, A = a)} dx \rightarrow 0, \quad (\text{S17})$$

as $K \rightarrow \infty$. Then, the nonparametric estimator $\hat{\tau}$ given in (14) is consistent for τ .

Proof S1 By Theorem S3,

$$\lim_{n \rightarrow \infty} \left\| \frac{\hat{f}(x \mid A = a, R = 1_p)}{\widehat{\text{pr}}(R = 1_p \mid x, A = a)} - \frac{f(x \mid A = a, R = 1_p)}{\text{pr}(R = 1_p \mid x, A = a)} \right\|_{\infty} = 0 \quad (\text{S18})$$

almost surely. Since $\hat{\tau}(x)$ and $\tau(x)$ are uniformly bounded in I_K for $K > B$, together with (S17) and (S18), for any ϵ , there exists $K_2 > 0$, such that for any $K > K_2$,

$$\lim_{n \rightarrow \infty} \text{pr} \left[\left| \int_{I_K^c} \hat{\tau}(x) \left\{ \frac{\hat{f}(x \mid A = a, R = 1_p)}{\widehat{\text{pr}}(R = 1_p \mid x, A = a)} - \frac{f(x \mid A = a, R = 1_p)}{\text{pr}(R = 1_p \mid x, A = a)} \right\} dx \right| > \frac{\epsilon}{4} \right] < \frac{\epsilon}{4}, \quad (\text{S19})$$

and

$$\lim_{n \rightarrow \infty} \left| \int_{I_K^c} \{\hat{\tau}(x) - \tau(x)\} \frac{f(x \mid A = a, R = 1_p)}{\text{pr}(R = 1_p \mid x, A = a)} dx \right| < \frac{\epsilon}{4}. \quad (\text{S20})$$

By Theorem S4, for any K ,

$$\lim_{n \rightarrow \infty} \left\| \hat{\tau}(x, R = 1_p) \frac{\hat{f}(x \mid A = a, R = 1_p)}{\widehat{\text{pr}}(R = 1_p \mid x, A = a)} - \tau(x, R = 1_p) \frac{f(x \mid A = a, R = 1_p)}{p(R = 1_p \mid x, A = a)} \right\|_{I_K} = 0 \quad (\text{S21})$$

almost surely, where $\|\cdot\|_I$ is defined in (S15). Therefore, for any ϵ , by (S21), we choose K_1

such that for any $K > K_1$,

$$\lim_{n \rightarrow \infty} \text{pr} \left\{ \left| \int_{I_{K_1}} \hat{\tau}(x, R = 1_p) \frac{\hat{f}(x | A = a, R = 1_p)}{\hat{\text{pr}}(R = 1_p | x, A = a)} dx - \int_{I_{K_1}} \tau(x, R = 1_p) \frac{f(x | A = a, R = 1_p)}{\text{pr}(R = 1_p | x, A = a)} dx \right| > \frac{\epsilon}{2} \right\} < \frac{\epsilon}{2}. \quad (\text{S22})$$

Combing (S19), (S20) and (S22), for any $\epsilon > 0$, we choose $K > \max(K_1, K_2)$,

$$\begin{aligned} & \lim_{n \rightarrow \infty} \text{pr}(|\hat{\tau} - \tau| > \epsilon) \\ = & \lim_{n \rightarrow \infty} \text{pr} \left\{ \left| \int \hat{\tau}(x) \frac{\hat{f}(x | A = a, R = 1_p)}{\hat{\text{pr}}(R = 1_p | x, A = a)} dx - \int \tau(x) \frac{f(x | A = a, R = 1_p)}{\text{pr}(R = 1_p | x, A = a)} dx \right| > \epsilon \right\} \\ \leq & \lim_{n \rightarrow \infty} \text{pr} \left\{ \left| \int_{I_K} \hat{\tau}(x) \frac{\hat{f}(x | A = a, R = 1_p)}{\hat{\text{pr}}(R = 1_p | x, A = a)} dx - \int_{I_K} \tau(x) \frac{f(x | A = a, R = 1_p)}{\text{pr}(R = 1_p | x, A = a)} dx \right| > \frac{\epsilon}{2} \right\} \\ + & \lim_{n \rightarrow \infty} \text{pr} \left[\left| \int_{I_K^c} \hat{\tau}(x) \left\{ \frac{\hat{f}(x | A = a, R = 1_p)}{\hat{\text{pr}}(R = 1_p | x, A = a)} - \frac{f(x | A = a, R = 1_p)}{\text{pr}(R = 1_p | x, A = a)} \right\} dx \right| > \frac{\epsilon}{4} \right] \\ + & \lim_{n \rightarrow \infty} \text{pr} \left[\left| \int_{I_K^c} \{\hat{\tau}(x) - \tau(x)\} \frac{f(x | A = a, R = 1_p)}{\text{pr}(R = 1_p | x, A = a)} dx \right| > \frac{\epsilon}{4} \right] < \epsilon, \end{aligned}$$

that is, $\hat{\tau}$ is consistent for τ .

For variance estimation of $\hat{\tau}$, a simple approach is to treat the nonparametric estimators as if they were parametric given the fixed tuning parameters, so that there is only a finite number of parameters. From this point of view, we can use typical approaches for variance estimation under parametric models. This approach has been shown to be asymptotically valid for nonparametric series regression; see, for example, Newey (1997). However, this variance estimation approach is not directly applicable in our context, since there is no simple analytical form for $\hat{\tau}$ in (13). That being said, in the light of treating the nonparametric estimators as if they were parametric, one might expect the nonparametric bootstrap to work for our estimator. For each bootstrap sample, we use the same tuning parameters, such as the smoothing parameter in the smoothing splines and the bandwidth in the kernel density

estimator, for all bootstrap samples. In our simulation study, the above bootstrap inference appears to be encouraging. Its theoretical properties are good topics for future research.

S9 Additional simulations

To assess the sensitivity of our nonparametric estimator to the choice of tuning parameters J and B , we specify a 4×3 design with $(J, B) \in \{(3, 50), (3, 100), (5, 50), (5, 100)\}$ and $n \in \{400, 800, 1600\}$. Table S3 shows the mean squared errors of the proposed estimator. For each choice of (J, B) , the mean squared error decreases with the sample size. The mean squared error decreases with J , and is relatively insensitive to the choice of B . The mean squared error remains small across all cases, which shows the promise of the proposed estimator.

Table S3: Simulation results: mean squared errors ($\times 10^{-3}$) of the proposed estimator of τ for different choices of (J, B) based on 2,000 Monte Carlo samples

(J, B)	$n = 400$	$n = 800$	$n = 1600$
(3, 50)	26.8	13.9	8.3
(3, 100)	27.0	14.1	8.7
(5, 50)	19.5	9.7	4.1
(5, 100)	21.3	10.2	4.5